Software Tools to Enable Reliable High-Performance Distributed Disk Arrays

Michael S. Warren
msw@lanl.gov

Ryan Joseph
rjoseph@t6.lanl.gov

M. Patrick Goda
pgoda@lanl.gov

http://space-simulator.lanl.gov

Los Alamos NATIONAL LABORATORY

Abstract

It is currently possible to construct a single-node RAID storage system with a 6 Terabyte capacity using commodity serial-ATA hard disk drives for less than $7000. Within the next year, a cluster of such systems (a distributed disk array) will be able to provide over a petabyte of storage for less than 1 million dollars. Obtaining reliablity and good performance from such a system is the focus of our project.

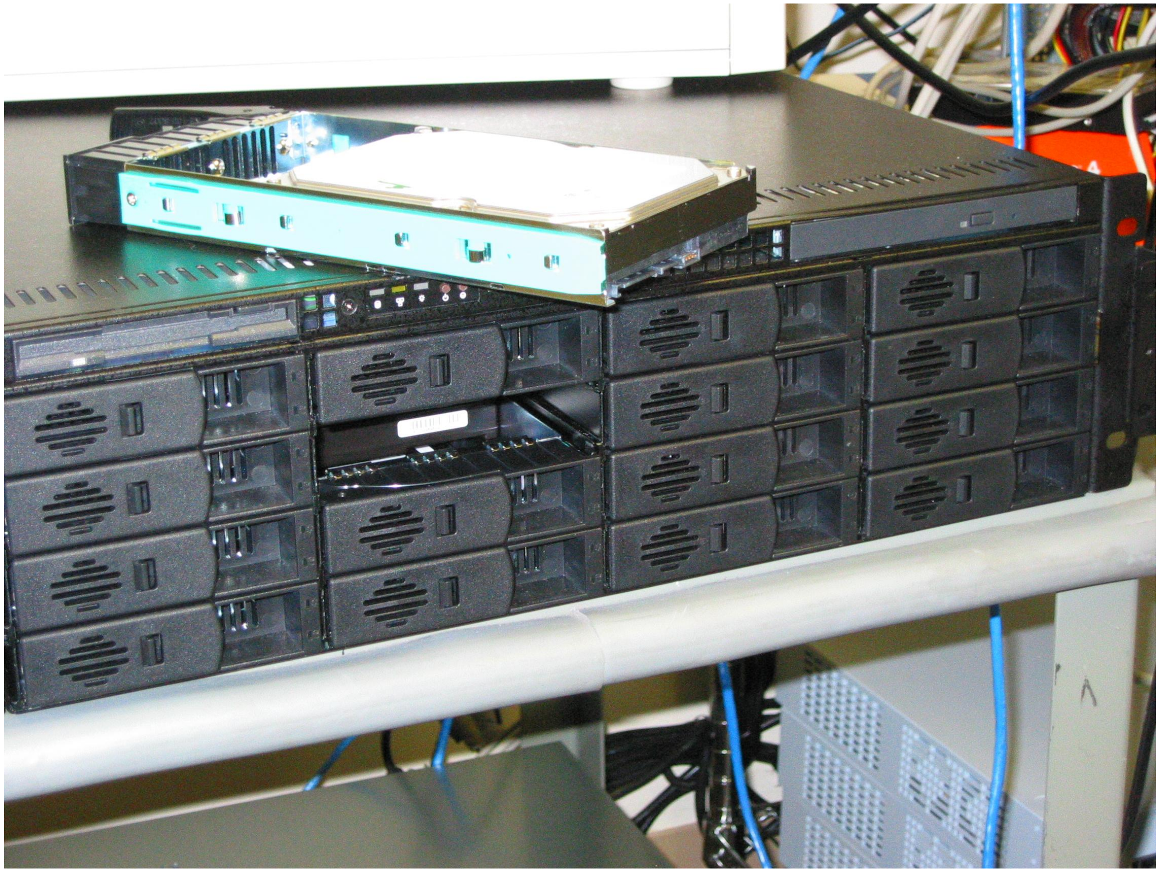
Over the course of this project, we have established a number of testbed systems containing over 100 Terabytes of storage. Significant progress to date includes detailed failure statistics on a variety of disk drives, performance benchmarks on a number of different systems, and modifications to the Linux Network Block Device (NBD) driver to support RAID-5 arrays across multiple cluster nodes.

Motivation

- The advent of commodity microprocessors with adequate floating-point performance and low-priced fast ethernet switches contributed to the emergence of Beowulf clusters in the mid-90s. We are currently poised for a similar advance in distributed disk arrays (DDAs), due to the dramatic decline in the price of commodity disk drives.

- The cost per Gbyte for 7200 RPM SATA disk drives is currently less than $1.00. Several groups have demonstrated fault-tolerant Terabyte RAID servers for a total cost of near $1000 per Terabyte. Used in a parallel cluster environment, multi-terabyte disk arrays with achievable read/write bandwidths that greatly exceed available Gigabit local and wide-area networking technology are possible. Additionally, the greater CPU/storage ratio in a DDA offers techniques which are not possible in traditional RAID arrays.

SATA Rack Mount

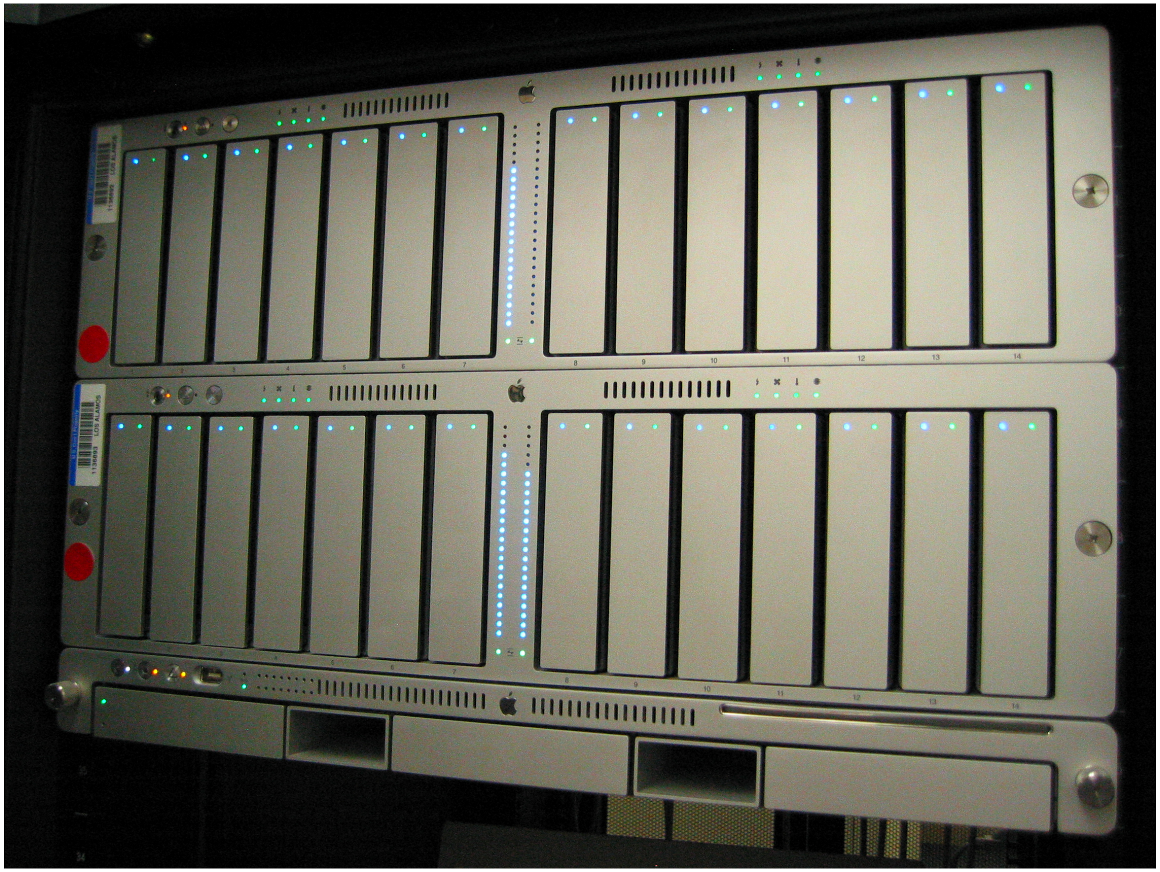


SATA Rack Mount in Production Use

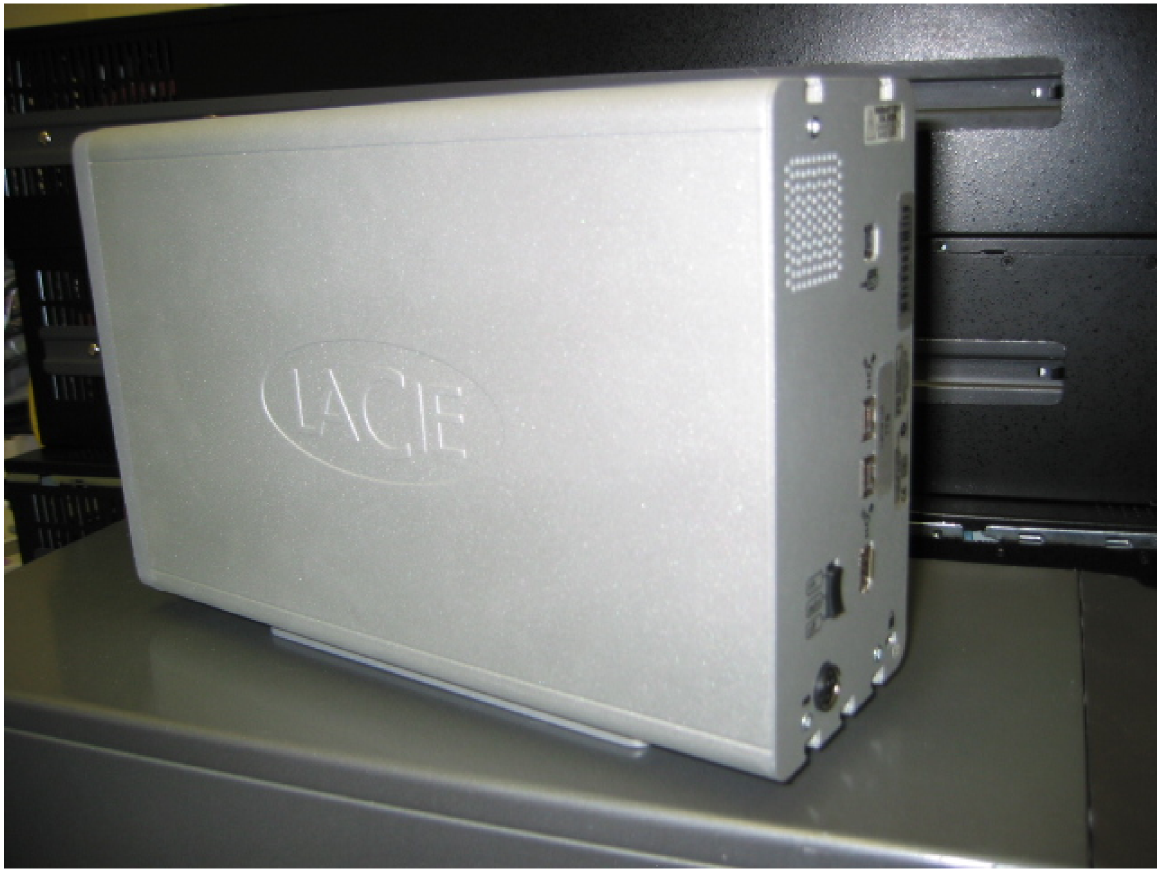
SATA Mid-Tower

SATA Mid-Towers in Production Use

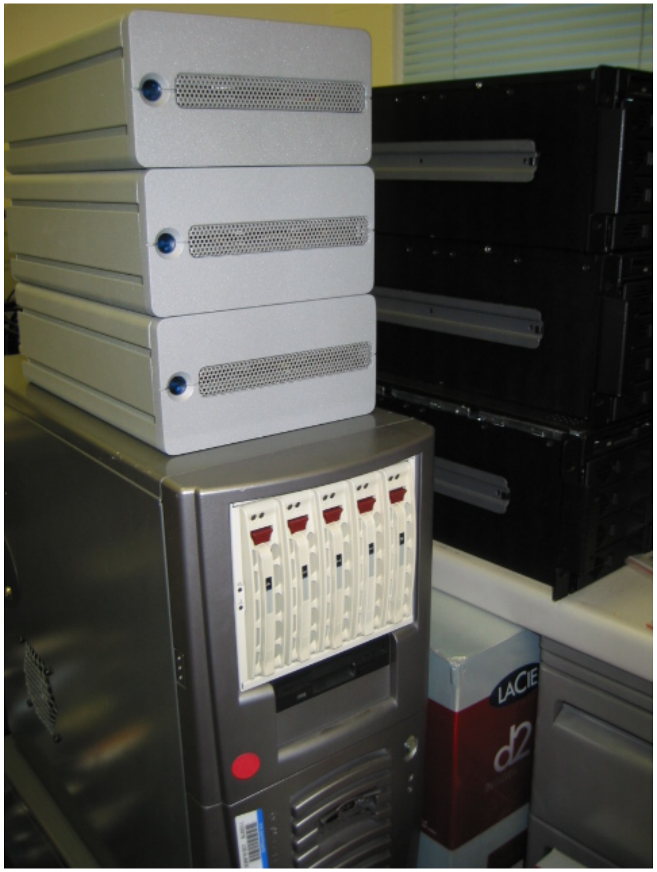
Xserve RAID



Lacie D2

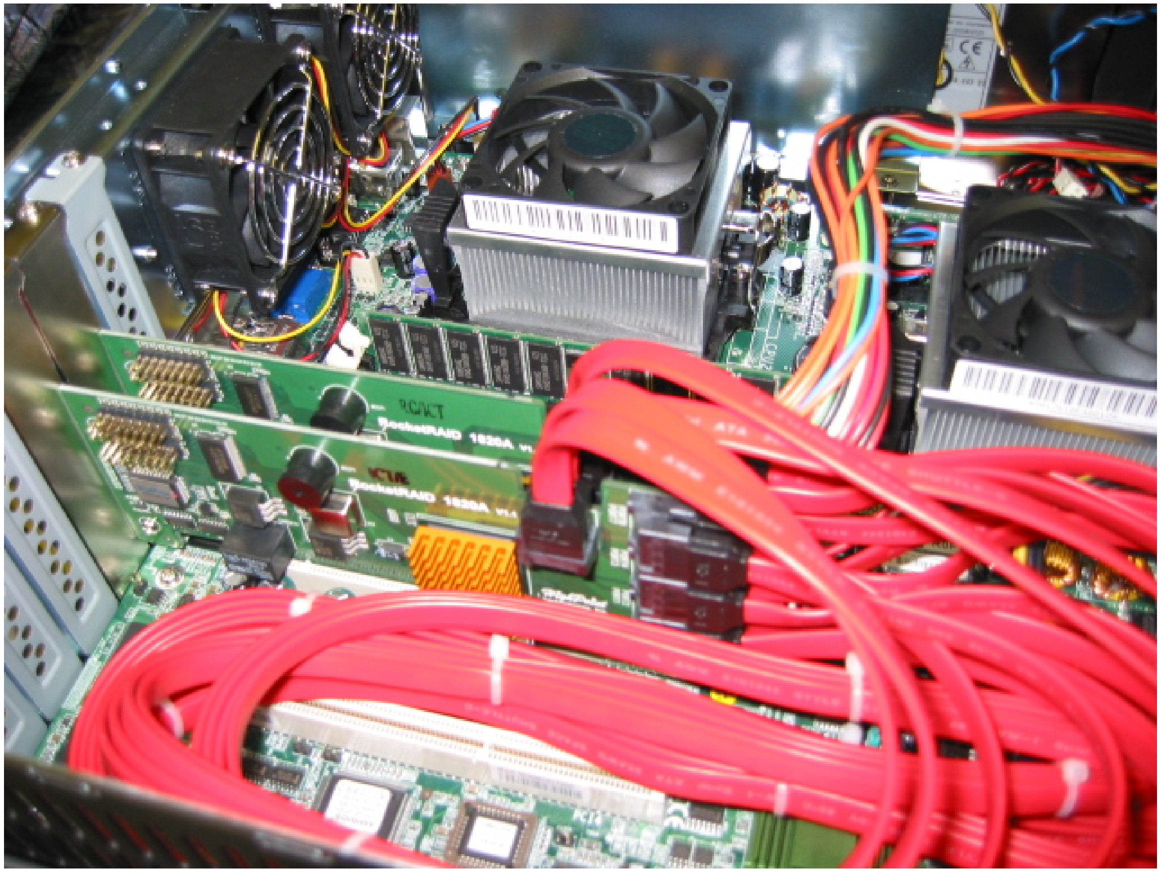


Lacie D2 RAID?



Latest Disk Technology

Latest RAID Card Technology



2 TB SATA Mid-Tower System

Qty.	Price	Ext.	Description
1	260	260	Intel P4 Processor 2.8GHz, 533MHz FSB
8	307	2456	Western Digital WD2500JD 250GB SATA 7200RPM
1	530	530	3ware 8506-8 RAID card
1	322	322	TYAN GC-SL S2707G2N-533 with Dual GigE and PCI-X
2	120	240	Corsair 2x512MB DDR266 PC2100 ECC Registered memory
1	150	150	Supermicro 5 Bay Hot-Swapable SATA HDD Enclosure
1	75	75	Mid-Tower case, 4x5.25 exposed, 4x3.5 int., extra fans
1	94	94	Enermax EG465P-VE (FCA) 431W Power Supply
1	80	80	Assembly
Total		$4207	$2.10 per Gbyte

Table 1: Mid-tower storage system pricing, August 2003

6 TB SATA 3U Rackmount System

Qty.	Price	Ext.	Description
1	180	180	Intel P4 Processor 3.0GHz, 800MHz FSB
15	315	4725	Hitachi Deskstar 400GB SATA 7200RPM
2	216	432	Rocketraid 1820A RAID card
1	230	230	TYAN Tomcat i7220 with Dual GigE and PCI-X
2	90	180	Corsair 512MB DDR400 PC3200 memory
1	854	854	Supermicro CSE-933T-R760B 3U 15-bay SATA chassis
1	80	80	Assembly
Total		$6681	$1.11 per Gbyte

Table 2: 3U Rackmount storage system pricing, April 2005

The Network Block Device

This project depends fundamentally on scaling disk storage from a single system to a parallel system. The networking abstraction used to communicate between systems is an important facet of this approach.

Recently, the iSCSI specification was developed in order to standardize communication with network attached SCSI devices. We investigated all of the major iSCSI implementations to see if they would provide a solid foundation to proceed from. The conclusions from this research were that iSCSI was over-engineered, and no robust implementations currently exist.

However, the “Network Block Device” exists as a standard part of the Linux kernel, and provides the functionality required for network attached storage. An NBD is ”a long pair of wires”. It makes a remote disk on a different machine act as though it were a local disk on your machine.

Challenges/Opportunities

Several interesting subtleties exist in the interaction between the block layer and the network layer in the Linux kernel. In particular, when the system becomes short on memory, dirty pages must be written to their backing store. However, if the act of writing those pages requires a memory allocation (as it normally would when writing to a TCP socket) the system will deadlock.

This problem is common to a number of “storage area network” applications. A number of solutions have been proposed, but as yet no consensus has been reached on the correct solution to the problem.

RAID Benchmarks

	Write	Re-Write	Read	Re-Read
3ware Hardware RAID-0	27.44	20.09	62.73	61.24
3ware Hardware RAID-5	27.44	20.09	62.73	61.24
3ware Software RAID-5	95.99	81.28	230.18	226.63
NBD baseline, 1kb BS	20.76	19.14	78.63	79.19
14-disk NBD RAID-5	91.23	93.46	56.38	56.78
tmpfs RAID-0	107.96	125.31	95.78	96.77
tmpfs RAID-5	81.07	87.52	90.86	90.90
3ware Software RAID-0	282.06	160.19	389.53	390.14
3ware Hardware RAID-5	161.26	98.77	254.21	257.66
3ware Software RAID-6	81.37	73.36	176.58	174.71
Highpoint Software RAID-0	353.18	247.63	443.32	445.43
Highpoint Software RAID-5	147.90	92.31	207.90	208.21
Apple Xserve RAID-5	43.02	40.73	133.77	147.33

Table 3: Hardware Storage Benchmarks (all data in Mbytes/sec)

Applications

- Warren is involved with the new Thinking Telescopes project at Los Alamos ($1M/yr for three years) which is depending on distributed disk arrays to enable transient detection analysis from the RAPTOR series of telescopes. The interface requirements of that system will further clarify the required performance and functionality of the DDA system.

- LANL will become a member institution of SDSS-II. We are investigating the possibility of hosting a data archive of the entire SDSS at Los Alamos.

- Warren is a member of the Destiny science team, and is involved in designing the data architecture for the Destiny JDEM concept study.

- We are looking for opportunities to collaborate with other proposed data-intensive missions such as LSST.